The Bell System Technical Journal
Vol. 52, No. 8, October, 1973
Printed in U.S.A.

Use of a Gate to Reduce the Variance of Delays in Queues With Random Service

By R. D. COLEMAN

(Manuscript received April 18, 1973)

We consider an N-server queuing system with Poisson arrivals and exponential service, in which arriving customers must pass through a gate into a waiting room before becoming eligible for service. Customers who find the gate closed wait outside until the gate opens; customers inside the waiting room are served at random. When the last customer inside acquires a server, the gate admits all those outside and then closes again. If no customer is waiting outside when the gate opens, the gate remains open until there is a queue of k waiting customers.

Service offered by this system is intermediary between random service and order-of-arrival service. As long as the gate is open and fewer than $N + k$ customers are in the system, service is purely random. The parameter k can be regarded as a threshold at which the queue is judged too long to permit random service to continue.

Our main results are (i) the Laplace-Stieltjes transform of the equilibrium distribution of the waiting time of an arbitrary customer and (ii) a comparison of the second moments of the waiting time for different values of k with those of the waiting time under random service and order-of-arrival service. The service is shown to be "nearly random" at low loads and "not quite order-of-arrival" at high loads; for higher values of k this transition occurs at higher traffic intensities.

I. INTRODUCTION

Switching systems, particularly electromechanical switching systems, are often constructed so that if several customers are awaiting service simultaneously, they will receive service in what is essentially random order, i.e., a server which becomes idle will choose its next customer at random from the queue of customers awaiting service. Such an arrangement may be satisfactory when the traffic intensity is low, but as the intensity increases, a progressively greater number of

customers will have to wait an undesirably long time, and the quality of service may become unacceptable.

There are available, however, methods of providing service other than "random service"; the most obvious one is service in order of arrival. The quality of service, which depends in part on the variance of the waiting time, will still diminisb as the traffic intensity increases, but not as quickly as when random service is used. In fact, order-of-arrival service is the "best" discipline (at least when the order of service does not depend on individual service times) in the sense that, for a given traffic intensity, and hence mean waiting time, the variance of the waiting time is smallest.[1] Unfortunately, it may not be worthwhile (or even possible) to build a system which offers service in order of arrival. One is therefore led to consider an intermediary queue discipline, one for which the variance of the equilibrium waiting time lies between that of random service and that of order-of-arrival service.

Suppose we have an M/M/N queuing system with customers arriving at rate λ and requiring a mean service time μ^{-1}. Suppose further that customers must first pass through a gate into a "corral," or "waiting room," before becoming eligible for service. So long as tbere are not more than N customers in the system, the gate remains open; an arriving customer enters the corral immediately and, if some server is idle, begins service. As soon as a customer has to wait (having found N customers in the system), the gate closes until that customer enters service. The gate then opens to admit all those who have accumulated outside the corral, and immediately closes again, remaining closed until all those inside have acquired a server, and so on. Tbus the customers arc admitted in bunchbes to tbe corral, and once inside, tbey are served at random. It is assumed tbat tbe corral bas an unlimited capacity. If there is no customer waiting when the gate opens, tbe gate remains open until tbere is again a customer who has to wait.

A gated queuing system merits consideration, not only because of the anticipated improvement over random service, but also because it can overcome design problems in certain telephone equipment which migbt otberwise lead to poor service for some customers. For example, in some switching equipment, eacb server hunts in a fixed sequence over a group of customer-sources; when a customer is found, tbe server stops to provide service and then resumes hunting from tbat point. Sucb a bunting procedure may be desirable from a hardware viewpoint, but it can result in unequal service among customers.

Gates have been successfully used to improve service in the manner described above, that is, by temporarily blocking subsequent bids for service until all those customers present have been served.

Some discussion of the gated queuing system is given in a 1953 paper by Wilkinson.[2] Theoretical results were summarized in 1958 by Beckwith,[3] although he gives few details as to how the results are derived. The model we consider in this paper is more general than the one described above. We shall assume that as soon as there are k customers (rather than one) waiting, the gate closes until all k customers have entered service. Thus the parameter k can be thought of as a threshold at which the queue is judged to be too long to permit purely random service to continue; the system enters the "gating mode," and the gate admits customers in bunches as described previously. If the gate opens and there is no customer waiting outside, then the system leaves the gating mode and the gate remains open until there is again a queue of k waiting customers.

In Section II we obtain a recurrence relation for the probability-generating function of the nth bunch-size and, after that, the generating function and moments of the equilibrium bunch-size distribution. Using these results we determine in Section III the Laplace-Stieltjes transform of the distribution of the equilibrium waiting time of an arbitrary customer. In Section IV we obtain the first two moments of the equilibrium waiting time, and we make some comparisons of the second moments of gated service for different values of the parameter k with those of random service and order-of-arrival service.

In several places in the text, we specialize a general result to the case $k = 1$, since that is the simplest of our gated systems. This allows us to verify that our results then agree with Beckwith's, and it often reduces a complicated expression to one that is more easily comprehended.

We can assume, without loss of generality, that the system is initially empty. Since we assume that the traffic intensity, $\rho = \lambda/N\mu$, is less than unity, the number of customers will always return to zero in a finite length of time, regardless of how many customers may be present at the beginning. Consequently, our equilibrium results do not depend on the initial conditions.

Several other authors have considered systems which operate in a manner similar to ours, but they all take $k = 1$ and assume that the underlying system is an M/G/1 queue. The "generations" in the M/G/1 queue, as described by Kendall[4] and later studied by Neuts,[5] correspond to the bunches in our model. Nair and Neuts[6,7] subse-

quently consider the waiting time distribution for the M/G/1 queue under the assumption that the queue discipline was either longest-processing-time-first or shortest-processing-time-first.

II. DISTRIBUTION OF THE EQUILIBRIUM BUNCH-SIZE

Let X_n be the number of customers in the nth bunch; that is, there are $N + X_n$ customers in the system the instant after the gate closes for the nth time. Thus X_n is the number of customers waiting outside if that number is positive, and is k if there was no one waiting, when the gate was opened for the $(n-1)$st time. Let T_n be the time it takes to serve X_n customers. Then, denoting by $p^d(T_n = t)$ the density of T_n at t, we can express the distribution of X_n in terms of that of X_{n-1} by

$$P(X_n = j) = \sum_{i=1}^{\infty} P(X_{n-1} = i) \int_0^{\infty} P(X_n = j \mid X_{n-1} = i,\ T_{n-1} = t) \cdot p^d(T_{n-1} = t \mid X_{n-1} = i)dt. \quad (1)$$

But, given that $X_{n-1} = i$, T_{n-1} has a gamma distribution with parameters i and $N\mu$; and

$$P(X_n = j \mid X_{n-1} = i,\ T_{n-1} = t) = P(X_n = j \mid T_{n-1} = t) = \begin{cases} \dfrac{(\lambda t)^j}{j!} e^{-\lambda t}, & j \neq k \\ \dfrac{(\lambda t)^k}{k!} e^{-\lambda t} + e^{-\lambda t}, & j = k. \end{cases}$$

The extra term $e^{-\lambda t}$ in this expression is the probability that no one is waiting when the gate opens; when this happens, the next bunch-size is necessarily k. Substituting these expressions into eq. (1) gives

$$P(X_n = j) = \sum_{i=1}^{\infty} P(X_{n-1} = i) \left(\frac{1}{1+\rho}\right)^i \times \left[\left(\frac{\rho}{1+\rho}\right)^j \binom{j+i-1}{j} + \delta_{jk}\right],$$

where $\rho = \lambda/N\mu$ and δ_{jk} is the Kronecker delta. Now, letting

$$P_n(u) = \sum_{j=1}^{\infty} P(X_n = j)u^j$$

be the probability-generating function of the distribution of X_n, we

have

$$P_n(u) = P_{n-1}\left(\frac{1}{1+\rho(1-u)}\right) + (u^k - 1)P_{n-1}\left(\frac{1}{1+\rho}\right). \quad (2)$$

Starting with $P_1(u) = u^k$, we can determine successively the $P_n(u)$. When $k = 1$, eq. (2) agrees with Beckwith's eq. (1).

We wish to obtain the distribution of the equilibrium bunch-size $X = \lim_{n\to\infty} X_n$. To see that an equilibrium distribution exists, we observe that the bunch-sizes X_n form a Markov chain which is irreducible and aperiodic. Since we are assuming $\rho < 1$, the number of customers in the system cannot grow without bound, so the states are not all transient or null. Therefore all states are ergodic, and there is a unique equilibrium distribution.[8] The distribution of $X = \lim_{n\to\infty} X_n$ has a probability-generating function $P(u) = \lim_{n\to\infty} P_n(u)$, which satisfies

$$P(u) = P\left(\frac{1}{1+\rho(1-u)}\right) + (u^k - 1)P\left(\frac{1}{1+\rho}\right).$$

This can be written in the form

$$P[r(u)] - P(u) = F(u)$$

by setting

$$r(u) = 1/[1 + \rho(1-u)]$$

and

$$F(u) = (1-u^k)P[1/(1+\rho)].$$

The solution to this functional equation is[9]

$$P(u) = \eta - \sum_{n=0}^{\infty} F[r_n(u)],$$

where r_n is the nth iterate of r and η is a constant. To evaluate this expression, we first need to find $r_n(u)$. We have $r_0(u) = u$ and

$$r_n = \frac{1}{1+\rho-\rho r_{n-1}}. \quad (3)$$

Letting y_n be such that

$$r_n = \frac{y_{n+1}}{y_n} + 1 + \frac{1}{\rho}$$

transforms eq. (3) into a linear homogeneous difference equation, which is easily solved. Thus we determine that

$$r_n(u) = \frac{1 - \rho u + (u-1)\rho^n}{1 - \rho u + (u-1)\rho^{n+1}}.$$

The solution to our functional equation is therefore

$$P(u) = \eta - \sum_{n=0}^{\infty} \left\{ 1 - \left[\frac{1 - \rho u + (u-1)\rho^n}{1 - \rho u + (u-1)\rho^{n+1}} \right]^k \right\} P\left(\frac{1}{1+\rho} \right).$$

Since $P(1) = 1$, we must have $\eta = 1$; since no bunch-size can be zero, $P(0) = 0$. This determines $P[1/(1+\rho)]$, so that finally

$$\begin{aligned} P(u) &= 1 - \frac{\sum_{n=0}^{\infty} \left\{ 1 - \left[\frac{1 - \rho u + (u-1)\rho^n}{1 - \rho u + (u-1)\rho^{n+1}} \right]^k \right\}}{\sum_{n=0}^{\infty} \left\{ 1 - \left[\frac{1-\rho^n}{1-\rho^{n+1}} \right]^k \right\}} \\ &= 1 - \frac{h(u)}{h(0)}, \end{aligned} \tag{4}$$

where $h(u)$ is the numerator in the second term of $P(u)$ above. By

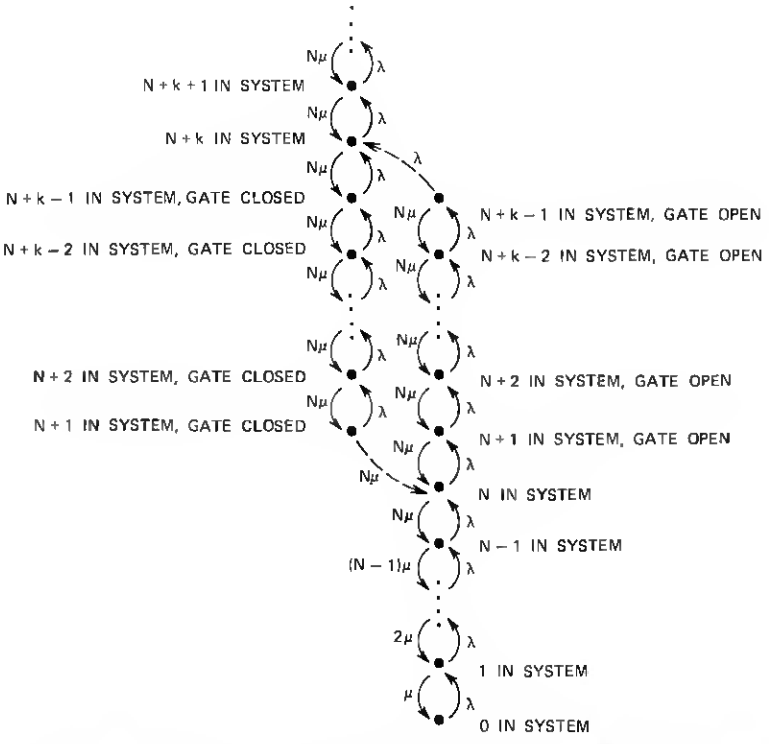

Fig. 1—Possible states of the system and the transition rates between states.

setting $k = 1$, we can reduce eq. (4) to Beckwith's result, his eq. (2):

$$P(u) = \frac{1}{g(1)}\left[\frac{u-1}{1-\rho} + g(1) - g\left(\frac{1-u}{\frac{1}{\rho}-u}\right)\right],$$

where

$$g(u) = u\sum_{n=0}^{\infty}\frac{\rho^n}{1-\rho^{n+1}u}.$$

The moments of X are obtained by differentiating $P(u)$. We find that

$$E(X) = P'(1) = \frac{k}{(1-\rho)h(0)}$$

and

$$\text{Var } (X) = \frac{k}{h(0)(1-\rho)}\left[\frac{k}{1+\rho} + \frac{\rho}{1-\rho} - \frac{k}{h(0)(1-\rho)}\right].$$

III. DISTRIBUTION OF THE EQUILIBRIUM WAITING TIME

Let W be the waiting time to the point of entering service of an arbitrary customer when the system is in equilibrium. The distribution of W depends on which state the system is in when the customer arrives; i.e., it depends on the number of customers already in the system and on whether the gate is open or closed. These states, together with the transition rates from one state to another, have been enumerated in Fig. 1. Let

$p_j = P[j \text{ customers in the system}], \qquad j \geqq 0,$

$p_j^c = P[j \text{ customers in the system, gate closed}],$
$\qquad j = N+1, N+2, \cdots, N+k-1,$

$p_j^o = P[j \text{ customers in the system, gate open}],$
$\qquad j = N+1, N+2, \cdots, N+k-1.$

Obviously, $p_j^c + p_j^o = p_j$. It is also clear that when $j \leqq N$, the gate is open, and that when $j \geqq N+k$, the gate is closed. The values of the p_j are just the equilibrium state probabilities of an M/M/N queue, which are

$$p_j = \frac{(\lambda/\mu)^j}{j!}p_0 = \frac{(\rho N)^j}{j!}p_0, \qquad j = 0, 1, \cdots, N$$

$$p_{N+j} = \rho p_{N+j-1} = \rho^j p_N = \rho^j\frac{(\rho N)^N}{N!}p_0, \qquad j > 0$$

$$p_0 = \left[\sum_{j=0}^{N-1}\frac{(\rho N)^j}{j!} + \frac{(\rho N)^N}{N!(1-\rho)}\right]^{-1}.$$

To find p_j^o, we equate the rates at which the system leaves and enters the state {j customers in the system, gate open}. From Fig. 1, we see that

$$(1 + \rho)p_{N+j}^o = \rho p_{N+j-1}^o + p_{N+j+1}^o, \qquad j = 1, 2, \cdots, k-1,$$

with $p_{N+k}^o = 0$ and $p_N^o = p_N$, where p_N is known from the above. The solution to this equation is

$$p_{N+j}^o = \frac{\rho^j - \rho^k}{1 - \rho^k} p_N, \qquad j = 0, 1, \cdots, k-1.$$

It then follows that

$$\begin{aligned} p_{N+j}^c &= p_{N+j} - p_{N+j}^o \\ &= \frac{1 - \rho^j}{1 - \rho^k} \rho^k p_N, \qquad j = 1, 2, \cdots, k-1. \end{aligned}$$

To find the distribution of the waiting time W, we shall consider what happens when an arriving customer encounters one of the following conditions:

H_1: $< N$ customers in the system,
H_2: the gate is closed,
$H_{3,j}$: $N + j$ customers in the system, and the gate is open, $j = 0, 1, \cdots, k-1$.

From the above computations,

$$P(H_1) = 1 - \sum_{j=0}^{\infty} \rho^j p_N = 1 - \frac{1}{1 - \rho} p_N, \tag{5}$$

$$P(H_2) = \sum_{j=1}^{k-1} p_{N+j}^c + \sum_{j=k}^{\infty} p_{N+j} = \frac{k\rho^k}{1 - \rho^k} p_N, \tag{6}$$

and

$$P(H_{3,j}) = \frac{\rho^j - \rho^k}{1 - \rho^k} p_N, \qquad j = 0, 1, \cdots, k-1. \tag{7}$$

An arriving customer who encounters H_1 immediately gains access to a server, so

$$P(W = t | H_1) = \begin{cases} 1, & t = 0 \\ 0, & t > 0 \end{cases}. \tag{8}$$

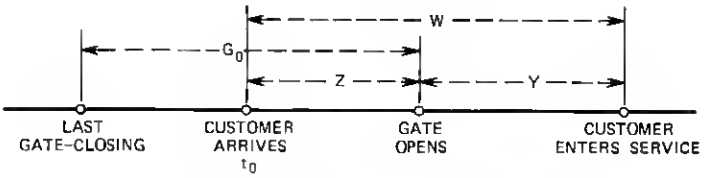

Fig. 2—Relations among the variables used.

An arriving customer who encounters H_2 will have to spend some time, Z, outside the gate waiting for the gate to open (see Fig. 2); once be gets inside, be will tben have to spend some additional time, Y, waiting to be chosen for service from the bunch he entered with. Tbe total amount of time tbe customer spends waiting to enter service is tberefore $W = Z + Y$, wberc Z can be regarded as the rcsidual lifetime of an interval G_0 during which the gate is closed. Thus we can write

$$p^d(W = t|H_2) = \int_{x=0}^{\infty}\int_{y=0}^{\infty} p^d(Z + Y = t,\, G_0 = y,\, Z = x)dydx$$

$$= \int_{x=0}^{\infty}\int_{y=0}^{\infty} p^d(Y = t - x|G_0 = y,\, Z = x)$$

$$\cdot p^d(Z = x|G_0 = y)p^d(G_0 = y)dydx. \quad (9)$$

We first need the distribution of G_0. Disregarding our arhitrary customer temporarily, let G be tbe equilibrium length of the time interval during which the gate is closed. If we are given that j customers entered the waiting room when tbe gate last opcned, then the gate will remain closed for j service times. Since the equilibrium probability that j customers entered when the gate last opened is $P(X = j)$, we bave

$$p^d(G = y) = \sum_{j=1}^{\infty} P(X = j)b^{*j}(y),$$

where $b(y) = N\mu \exp(-N\mu y)$ and the asterisk denotes convolution. The mean of this distribution is easily found to be

$$E(G) = \frac{E(X)}{N\mu} = \frac{k}{N\mu(1 - \rho)h(0)}.$$

Let t_0 he tbe epocb at which the arbitrary customer arrives, and let

G_0 be the length of the G-interval containing t_0. Then (see Ref. 10)

$$p^d(G_0 = y) = \frac{y}{E(G)} p^d(G = y)$$

$$= \frac{1}{k} N\mu y(1 - \rho)h(0) \sum_{j=1}^{\infty} P(X = j)b^{*j}(y). \quad (10)$$

Next, the distribution of Z, given $G_0 = y$, is uniform on $(0, y)$, so

$$p^d(Z = x | G_0 = y) = \begin{cases} \frac{1}{y}, & 0 \leqq x \leqq y \\ 0, & x > y. \end{cases} \quad (11)$$

We also have

$$p^d(Y = t - x | G_0 = y, Z = x) = p^d(Y = t - x | G_0 = y)$$

$$= \sum_{n=1}^{\infty} p^d(Y = t - x | n \text{ arrivals in } (0, y), G_0 = y)$$

$$\cdot p^d(n - 1 \text{ other arrivals in } (0, y) | G_0 = y)$$

$$= \sum_{n=1}^{\infty} \frac{1}{n} \left(\sum_{l=1}^{n} b^{*l}(t - x) \right) \cdot \frac{(\lambda y)^{n-1}}{(n - 1)!} e^{-\lambda y}, \quad (12)$$

where the first factor in each summand reflects the fact that the arbitrarily chosen customer has probability $1/n$ of waiting one service time, $1/n$ of waiting two service times, $\cdots$, and $1/n$ of waiting n service times. Substituting eqs. (10), (11), and (12) into eq. (9), we obtain

$$p^d(W = t | H_2)$$

$$= \frac{1}{k} N\mu(1 - \rho)h(0) \sum_{j=1}^{\infty} \sum_{n=1}^{\infty} P(X = j) \sum_{l=1}^{n} \int_{x=0}^{t} b^{*l}(t - x)$$

$$\cdot \int_{y=x}^{\infty} b^{*j}(y) e^{-\lambda y} \frac{(\lambda y)^{n-1}}{n!} dy dx. \quad (13)$$

Notice that the range of integration of the inner integral was reduced from $(0, \infty)$ to (x, ∞) because of eq. (11), and that of the outer integral was reduced from $(0, \infty)$ to $(0, t)$ because $b^{*l}(t - x) = 0$ for $x > t$. By setting $k = 1$, $N\mu = 1$, and $\lambda = \rho$ in eq. (13), we can obtain Beckwith's expression for the corresponding density in his model, the density of W given that the arbitrary customer finds more than N customers in the system.

We can calculate the Laplace-Stieltjes transform, $\psi(s)$, of the distribution (13), but the algebra is long and tedious. The results are

given below, one in terms of $P(X = j)$ and the other explicitly:

$$\psi(N\mu s)$$

$$= \frac{(1-\rho)h(0)}{\rho k s^2}\left\{P(X=1)\log\frac{(1+s)(1+s+s\rho)}{(1+s+\rho)(1+s)-\rho}\right.$$

$$+\sum_{j=2}^{\infty}\frac{1}{j-1}P(X=j)\left[1-\left(\frac{1}{1+s}\right)^{j-1}-\left(\frac{1+s}{1+s+s\rho}\right)^{j-1}\right.$$

$$\left.\left.+\left(\frac{1+s}{(1+s)(1+\rho+s)-\rho}\right)^{j-1}\right]\right\} \quad (14)$$

$$= \frac{1-\rho}{k\rho s^2}\sum_{n=0}^{\infty}\left[\frac{k\rho^n(1-\rho)^2}{(1-\rho^{n+1})^2}\left\{\left(\frac{1-\rho^n}{1-\rho^{n+1}}\right)^{k-1}\right.\right.$$

$$\cdot\log\left(\frac{(1+s-\rho-s\rho^{n+1})(1+s-\rho-s\rho^{n+2})}{(1-\rho)[(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}]}\right)$$

$$+\sum_{j=2}^{k}\binom{k-1}{j-1}\left(\frac{1-\rho^{n-1}}{1-\rho^n}\right)^{k-j}\left(\frac{\rho^{n-1}(1-\rho)^2}{(1-\rho^n)(1-\rho^{n+1})}\right)^{j-1}$$

$$\cdot\sum_{m=1}^{j-1}\frac{1}{m}\left[\left(\frac{1-\rho^{n+1}}{1-\rho}\right)^m-\left(\frac{(1-\rho^{n+1})(1+s)}{1+s-\rho-s\rho^{n+1}}\right)^m\right.$$

$$-\left(\frac{(1-\rho^{n+1})(1+s+s\rho)}{1+s-\rho-s\rho^{n+2}}\right)^m$$

$$\left.\left.+\left(\frac{(1-\rho^{n+1})[(1+s)^2+s\rho]}{(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}}\right)^m\right]\right\}$$

$$-\frac{1}{1-\rho^{n+1}}\left\{(1-\rho)-\frac{(1+s-\rho-s\rho^n)^k}{(1+s-\rho-s\rho^{n+1})^{k-1}}\right.$$

$$-\frac{1}{1+s}\cdot\frac{(1+s-\rho-s\rho^{n+1})^k}{(1+s-\rho-s\rho^{n+2})^{k-1}}$$

$$\left.+\frac{1}{1+s}\cdot\frac{[(1+s)^2-\rho-s(1+s+\rho)\rho^n]^k}{[(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}]^{k-1}}\right\}\Bigg]. \quad (15)$$

When k is set equal to 1 in eq. (15), the sum from 2 to k is trivially zero, and the expression in the last pair of braces collapses to zero; the transform then reduces to

$$\psi(N\mu s) = \frac{(1-\rho)^3}{\rho s^2}\sum_{n=0}^{\infty}\frac{\rho^n}{(1-\rho^{n+1})^2}$$

$$\times\log\left[\frac{(1+s-\rho-s\rho^{n+1})(1+s-\rho-s\rho^{n+2})}{(1-\rho)[(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}]}\right],$$

$$(k=1). \quad (16)$$

The final portion of the waiting time distribution needed is $f_j(t) = p^d(W = t | H_{3,j})$, $j = 0, 1, \cdots, k-1$. Let W_j be the waiting time of a customer who arrives to find the gate open and $N + j$ in the system. Then W_j has the density $f_j(t)$, and, letting $E_\lambda(t)$ denote the exponential density $\lambda e^{-\lambda t}$,

$$\begin{cases} f_j = \dfrac{\lambda}{\lambda + N\mu} E_{\lambda+N\mu} * f_{j+1} + \dfrac{N\mu}{\lambda + N\mu} \\ \qquad \cdot \left[\dfrac{1}{j+1} E_{\lambda+N\mu} + \dfrac{j}{j+1} E_{\lambda+N\mu} * f_{j-1} \right], \\ \qquad\qquad\qquad\qquad\qquad\qquad j = 0, 1, \cdots, k-2 \\ f_{k-1} = \dfrac{1}{k} E_{N\mu} + \dfrac{1}{k} E_{N\mu}^{*2} + \cdots + \dfrac{1}{k} E_{N\mu}^{*k}. \end{cases}$$

The reasoning behind these equations is that, if an arriving customer finds the gate open and $j (< k-1)$ customers waiting, then either the next event is an arrival, in which case he waits until that event plus an additional time distributed as W_{j+1}, or else the next event is a departure, in which case he waits until that event, after which either he is served immediately [with probability $1/(j+1)$] or someone else is chosen and our customer must wait an additional time distributed as W_{j-1} [with probability $j/(j+1)$]. If, on the other hand, the arriving customer finds the gate open and $k-1$ customers waiting, then the gate shuts behind him and he waits either $1, 2, \cdots$, or k service times, each having probability $1/k$. Taking Laplace-Stieltjes transforms of the equations, and denoting the transform of $f_j(t)$ by $\phi_j(s)$, we obtain

$$\begin{cases} \rho j \phi_j(N\mu s) = (1 + \rho + s) j \phi_{j-1}(N\mu s) - (j-1)\phi_{j-2}(N\mu s) - 1, \\ \qquad\qquad\qquad\qquad\qquad\qquad j = 1, \cdots, k-1, \\ \phi_{k-1}(N\mu s) = \dfrac{1}{ks} \left[1 - \left(\dfrac{1}{1+s} \right)^k \right]. \end{cases} \tag{17}$$

For any particular k, this set of equations can be solved explicitly by successive substitution. Finally, using eqs. (5), (6), (7), and (8), we can represent the Laplace-Stieltjes transform of the distribution of the waiting time W as

$$\phi(N\mu s) = 1 - \frac{1}{1-\rho} p_N + p_N \sum_{j=0}^{k-1} \frac{\rho^j - \rho^k}{1-\rho^k} \phi_j(N\mu s) + \frac{k\rho^k}{1-\rho^k} p_N \psi(N\mu s), \tag{18}$$

where $\phi_j(N\mu s)$ is given by eq. (17) and $\psi(N\mu s)$ is given by eq. (14)

or eq. (15). When $k = 1$, eq. (18), with the help of eq. (16), can be written explicitly as

$$\phi(N\mu s) = 1 - \frac{1}{1-\rho}p_N + \frac{1}{1+s}p_N + \frac{(1-\rho)^2}{s^2}p_N \sum_{n=0}^{\infty} \frac{\rho^n}{(1-\rho^{n+1})^2} \cdot \log\left[\frac{(1+s-\rho-s\rho^{n+1})(1+s-\rho-s\rho)^{n+2}}{(1-\rho)[(1+s)^2-\rho-s(1+s+\rho)\rho^{n+1}]}\right], \quad (k=1).$$

IV. MOMENTS OF THE EQUILIBRIUM WAITING TIME

Since the mean waiting time does not depend on the queue discipline (see Ref. 11), the mean is the same as for a simple queue with service in order of arrival, i.e.,

$$E(W) = \sum_{j=0}^{\infty} p_N \rho^j \frac{j+1}{N\mu} = \frac{p_N}{N\mu(1-\rho)^2}.$$

The second-moment computations are fairly lengthy, finally yielding

$$E(W^2) = \frac{2p_N}{N\mu(1-\rho^k)}[M'(\rho) + M(\rho) - \rho^{k-1}M'(1) - \rho^{k-1}M(1)] + \frac{kp_N\rho^{k-1}(1-\rho)}{(N\mu)^2(1-\rho^k)}\left\{\frac{(k-1)(k-2)}{6}\left[\frac{2+2\rho+3\rho^2}{1-\rho^3}\right] + (k-1)\left[\frac{2+3\rho+3\rho^2-\rho^4}{(1-\rho^2)(1-\rho^3)}\right] + \frac{2+4\rho+5\rho^2+2\rho^3+\rho^4}{(1-\rho)(1-\rho^2)(1-\rho^3)}\right\}, \quad (19)$$

where M is a function defined by

$$M(x) = -\sum_{j=0}^{k-1} x^j \phi_i'(0) = \sum_{j=0}^{k-1} x^j E(W \mid H_{3,j}).$$

The variance of W is obtained by subtracting the square of the mean of W:

$$\text{Var}(W) = E(W^2) - \frac{p_N^2}{(N\mu)^2(1-\rho)^4}.$$

For comparison purposes, we also need the second moments for order-of-arrival service and for random service. When service is in the order of arrival, we obtain[11]

$$E(W^2) = E(W^2 \mid W > 0)P(W > 0) = \frac{2p_N}{(N\mu)^2(1-\rho)^3}. \quad (20)$$

When service is at random, the second moment can be written as[11]

$$E(W^2) = \frac{2p_N}{(N\mu)^2(1-\rho)^3} \cdot \frac{2}{2-\rho}. \quad (21)$$

Observe that the second moments depend on N only through the factor $p_N/(N\mu)^2$, assuming the value of ρ is fixed. This is true also for the second moment in eq. (19), since each of the M-terms contains a factor $(N\mu)^{-1}$. It is therefore convenient to consider the ratio of $E(W^2)$ for the gated system [eq. (19)] to the second moment for the order-of-arrival system [eq. (20)], i.e.,

$$R(k, \rho) = \frac{N\mu(1-\rho)^3}{1-\rho^k}[\rho M'(\rho) + M(\rho) - \rho^{k-1}M'(1) - \rho^{k-1}M(1)]$$
$$+ \frac{k\rho^{k-1}(1-\rho)^4}{2(1-\rho^k)}\left\{\frac{(k-1)(k-2)}{6}\left[\frac{2+2\rho+3\rho^2}{1-\rho^3}\right]\right.$$
$$\left. + (k-1)\left[\frac{2+3\rho+3\rho^2-\rho^4}{(1-\rho^2)(1-\rho^3)}\right] + \frac{2+4\rho+5\rho^2+2\rho^3+\rho^4}{(1-\rho)(1-\rho^2)(1-\rho^3)}\right\}. \quad (22)$$

Because this ratio is independent of N, it provides a useful tool for examining the effect of the value of k on the second moment of the waiting time W. Thus we shall be interested in determining its properties.

The function $R(k, \rho)$ has been plotted in Fig. 3 on the interval $0 < \rho < 1$ for a variety of values of k. We observe that $R(k, \rho)$ is bounded from below by unity, increases as k increases, and is bounded from above by $2/(2-\rho)$, the ratio of eq. (21) and eq. (20), which corresponds to $k = \infty$. This general behavior is, of course, just what we expected *a priori*. In order to demonstrate analytically that

$$1 \leqq R(k, \rho) \leqq \frac{2}{2-\rho}, \quad (23)$$

we first introduce the inequality

$$\frac{1}{N\mu}\frac{j+2}{2} \leqq E(W \mid H_{3,j}) \leqq \frac{1}{N\mu}\frac{j+2}{2}\frac{2}{2-\rho},$$
$$j = 0, 1, \cdots, k-1. \quad (24)$$

The left half of this inequality is demonstrated by considering what happens when an arriving customer finds j others waiting, but no more arrivals are permitted to enter the system: the customer's expected waiting time will decrease to $(j+2)/2N\mu$, since the customer will have to wait either 1, 2, $\cdots$, or $j+1$ service times, each with probability $(j+1)^{-1}$. The right half of the inequality is demonstrated by considering what happens when there is no gate at all to block future arrivals when a threshold k is reached: the mean waiting time $E(W \mid H_{3,j})$ would then increase to the corresponding mean in a system

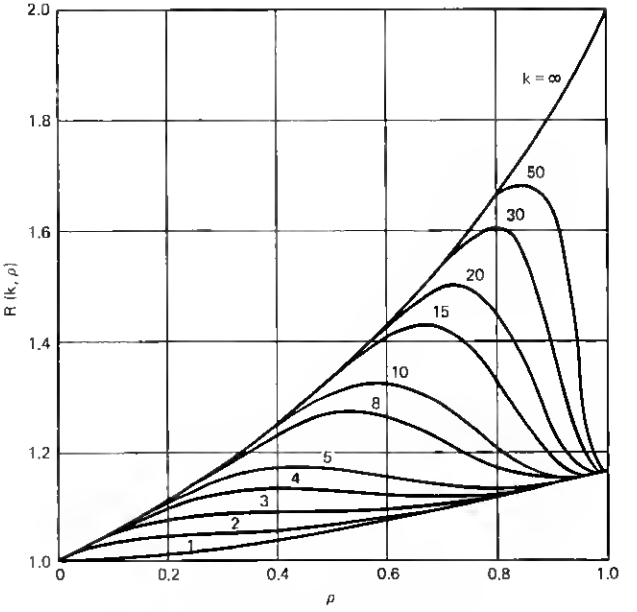

Fig. 3—Ratio of second moments.

employing purely random service. In a random service system, a customer who arrives to find j others waiting has an expected delay of

$$\frac{1}{N\mu}\cdot\frac{j+2}{2}\cdot\frac{2}{2-\rho}, \qquad j = 0, 1, \cdots, k-1,$$

a fact whicb is derived in the appendix.

Using the left half of eq. (24), we can substitute $(j+2)/2N\mu$ for $E(W \mid H_{3,j})$ in $R(k, \rho)$ and combine terms, obtaining

$$R(k, \rho) \geqq 1 + \frac{k\rho^{k+1}(1-\rho)}{12(1-\rho^k)}\left\{2 + \frac{3k(1-\rho)(1-\rho^2)}{(1+\rho)(1+\rho+\rho^2)} + \frac{k^2(1-\rho)^2}{1+\rho+\rho^2}\right\},$$

from which it is obvious that $R(k, \rho) \geqq 1$. Similarly, using the right half of eq. (24), we can substitute $(j+2)/N\mu(2-\rho)$ in $R(k, \rho)$. Combining terms, we obtain

$$R(k, \rho) \leqq \frac{2}{2-\rho} - \frac{k\rho^k(1-\rho)}{(2-\rho)(1-\rho^k)}\left\{\frac{k^2(1-\rho)^2(2+3\rho^2)}{12(1+\rho+\rho^2)} + \frac{k(1-\rho)(2+2\rho+5\rho^2+\rho^4)}{4(1+\rho)(1+\rho+\rho^2)} + \frac{2+7\rho+10\rho^2+8\rho^3+3\rho^4}{6(1+\rho)(1+\rho+\rho^2)}\right\},$$

from which it is clear that $R(k, \rho) \leqq 2/(2 - \rho)$. Thus, eq. (23) is established.

Perhaps the most striking feature of Fig. 3 is that all the curves approach the same value, 7/6, as $\rho \to 1$. It is easy to demonstrate, by using eq. (22), that this must occur. Since the conditional means $E(W \mid H_{3,j})$ remain bounded as $\rho \to 1$, $M(\rho) \to M(1)$ and both are finite. Thus the first term of $R(k, \rho)$ approaches zero as $\rho \to 1$. In the second term, the factor $(1 - \rho)^4$ in front causes all but the last term in braces to approach zero. Thus,

$$R(k, 1) = \lim_{\rho \to 1} \frac{k\rho^{k-1}}{2(1 + \rho + \cdots + \rho^{k-1})} \cdot \frac{2 + 4\rho + 5\rho^2 + 2\rho^3 + \rho^4}{(1 + \rho)(1 + \rho + \rho^2)} = \frac{7}{6}.$$

In order to gain some insight as to why the curves meet at a common point at $\rho = 1$, we will find it helpful to consider a supplementary variable, the fraction of time the system spends in the gating mode. When the system is in equilibrium, this is simply the probability that an arriving customer finds the gate closed, and is given by eq. (6):

$$P(\text{system is in gating mode}) = \frac{k\rho^k}{1 - \rho^k} p_N.$$

This quantity has been plotted in Fig. 4 as a function of ρ, for the arbitrarily chosen value $N = 7$. It can easily be seen (and is intuitively obvious) that when ρ is very close to 1, the system spends almost all its time in the gating mode. But when the system is in the gating mode, the system's operation is independent of the value of k; it is only when the gate is open that the threshold value k can have any effect. Thus, as ρ approaches 1, the system becomes independent of k; so we can expect the curves in Fig. 3 to be independent of k at $\rho = 1$.

Another feature of the curves in Fig. 3 is that the slope at zero for $k \geqq 2$ is the same as the slope of $2/(2 - \rho)$; but the slope for $k = 1$ is zero. The reason for this becomes clear when we realize that when $k = 1$, order-of-arrival service is guaranteed until there are $N + 3$ customers in the system, while $k \geqq 2$ makes it possible for passing to occur as soon as there are $N + 2$ customers in the system. For small ρ, $N + 3$ in the system is much less likely than $N + 2$.

The ratio in eq. (22) is convenient because it is independent of N. The variance of a distribution, however, is also frequently of interest. It is clear from Fig. 3 that the second moments, and therefore the variances, increase with increasing k. There is, therefore, a similar family of curves, one family for each value of N, obtained by computing the ratio of the variance with threshold k to the variance with order-of-

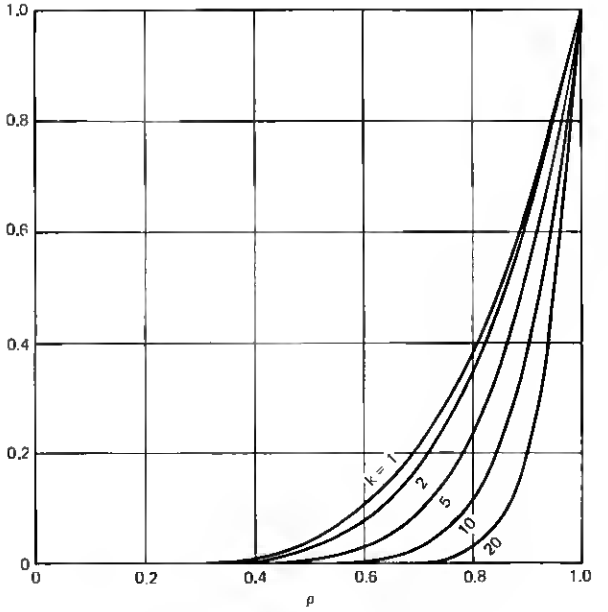

Fig. 4—Equilibrium probability that gate is closed ($N = 7$).

arrival service. In Fig. 5 we have plotted some of these curves, again for $N = 7$. It is obvious that these curves have the same general shape as those in Fig. 3. The main differences are (*i*) that the ratio of the variances when $k = \infty$ (random service) goes to 3 as $\rho \to 1$, while the ratio of the second moments when $k = \infty$ goes to 2, and (*ii*) that the ratios of the variances for $k < \infty$ go to 8/6 as $\rho \to 1$, while the ratios of the second moments go to 7/6. These facts are easily verified analytically by letting $\rho \to 1$ in the actual expressions for the ratios.

Since the second moments (and variances) increase as a function of k, there arises the question of why one might consider a threshold value k greater than 1. Clearly, if the queuing systems were otherwise equivalent, one would prefer $k = 1$ to any larger value of k. But it is equally possible that a queuing system would be more costly to operate when it is in the gating mode, since more bookkeeping is necessary: each waiting customer must be classified as "inside" or "outside" the waiting room, and these labels must be changed when the gate operates. One would then prefer a system in which the gate were used as little as possible (large k), consistent with an acceptable quality of service. The resulting tradeoff between cost and quality of service

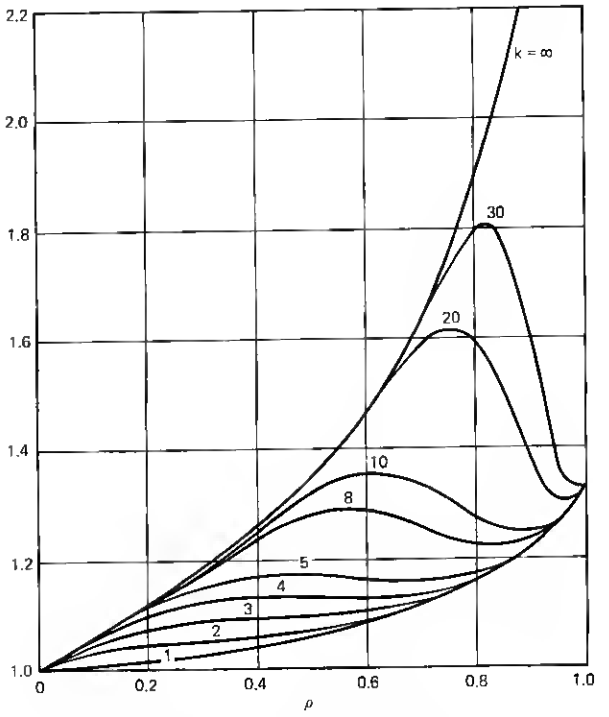

Fig. 5—Ratio of variances ($N = 7$).

can be resolved only by examination of the particular application at hand.

V. CONCLUDING REMARKS

In summary, our main result is the specification of the distribution of the equilibrium waiting time of an arbitrary customer in a queuing system whose service discipline is a compromise between service in order of arrival and random service. Our model contains a parameter, k, which determines how "close" the discipline is to order-of-arrival service or to random service. We have seen that the variance of the waiting time is bounded from below by the variance for order-of-arrival service, and that as k increases, the variance increases, approaching the variance of the waiting time when random service is employed. We also found a convenient quantity, $R(k, \rho)$, which is independent of the number of servers, and which, together with Fig. 3, allowed us to examine the effect of the threshold k on the waiting time.

Figure 3 shows how, for fixed $k > 1$, the service changes from "nearly random" to "not quite order-of-arrival" with increasing load, and how this transition occurs at higher loads as k increases.

There are, of course, other variables which can be derived from the system we have described; for example, in studies of equipment life, it might be useful to know the distribution of the number of gate-closings that occur in an interval of length t. It did not seem worthwhile to pursue such questions in the present study, which deals with gating from the viewpoint of traffic performance.

APPENDIX

Suppose we have an M/M/N queuing system employing random service, with arrival rate λ and mean holding time μ^{-1}. The mean waiting time to the point of entering service of an arbitrary customer in such a system is

$$m = \frac{p_N}{N\mu(1-\rho)^2},$$

where $\rho = \lambda/N\mu < 1$ and p_N is the known equilibrium probability that there are exactly N customers in the system. We wish to determine the mean waiting time, m_j, of a customer who arrives to find $N + j$ other customers in the system. The m_j's satisfy

$$m_j = \frac{1}{\lambda + N\mu} + \frac{\lambda}{\lambda + N\mu} m_{j+1} + \frac{N\mu}{\lambda + N\mu} \cdot \frac{j}{j+1} m_{j-1}, \quad j \geqq 0. \quad (25)$$

The rationale for this equation is that a customer must wait at least until the next change of state; the mean of this initial delay is $(\lambda + N\mu)^{-1}$. If the change of state is caused by an arrival [which occurs with probability $\lambda/(\lambda + N\mu)$], then the customer will have to wait an additional period of time whose mean is m_{j+1}. If, on the other hand, the change of state is caused by a departure [which occurs with probability $N\mu/(\lambda + N\mu)$], then with probability $j/(j+1)$ our customer will not be chosen from the group of $j + 1$ customers, and he will have to wait an additional period of time whose mean is m_{j-1}.

We now introduce the function

$$H(x) = \sum_{j=0}^{\infty} m_j x^j.$$

This series converges for $x \leqq \rho$, since the mean waiting time of an arbitrary customer is

$$m = p_N H(\rho) = \frac{p_N}{N\mu(1-\rho)^2}. \quad (26)$$

Multiplying eq. (25) by $(1 + \rho)(j + 1)x^j$ and summing on j, we can obtain a first-order differential equation:

$$H'(x) + \frac{1 + \rho - x}{(1 - x)(x - \rho)} H(x) = \frac{1}{N\mu} \frac{1}{(1 - x)^3(x - \rho)}.$$

The solution to this equation is

$$H(x) = C(1 - x)^\rho \left(\frac{1 - x}{x - \rho}\right)^{1/(1-\rho)} + \frac{2 - x}{N\mu(1 - x)^2(2 - \rho)}.$$

Using eq. (26) as the boundary condition, we see that C must be zero in order that $H(x)$ remain finite as $x \to \rho$. Thus

$$H(x) = \frac{2 - x}{N\mu(1 - x)^2(2 - \rho)}.$$

The power series expansion of this function is found to be

$$H(x) = \sum_{j=0}^{\infty} \frac{j + 2}{N\mu(2 - \rho)} x^j;$$

therefore, the means we desire are given by

$$m_j = \frac{j + 2}{N\mu(2 - \rho)}.$$

REFERENCES

1. Kingman, J. F. C., "The Effect of Queue Discipline on Waiting Time Variance," Proc. Cambridge Phil. Soc., *58* (1962), pp. 163–164.
2. Wilkinson, R. I., "Working Curves for Delayed Exponential Calls Served in Random Order," B.S.T.J., *32*, No. 2 (March 1953), pp. 360–383.
3. Beckwith, D. A., "Delay at a Simple Gate," unpublished work, May 16, 1958.
4. Kendall, D. G., "Some Problems in the Theory of Queues," J. Roy. Stat. Soc., Series B, *13*, No. 2 (1951), pp. 151–185.
5. Neuts, M. F., "The Queue with Poisson Input and General Service Times, Treated as a Branching Process," Duke Math. J., *36*, No. 2 (June 1969), pp. 215–231.
6. Nair, S. S., and Neuts, M. F., "A Priority Rule Based on the Ranking of the Service Times for the M/G/1 Queue," Operations Res., *17*, No. 3 (May-June 1969), pp. 466–477.
7. Nair, S. S., and Neuts, M. F., "An Exact Comparison of the Waiting Times Under Three Priority Rules," Operations Res., *19*, No. 2 (March-April 1971), pp. 414–423.
8. Feller, W., *An Introduction to Probability Theory and its Applications, Volume I*, New York: John Wiley and Sons, 1957, p. 355.
9. Kuczma, M., *Functional Equations in a Single Variable*, New York: Hafner Publishing Company, 1968, pp. 46–58.
10. Feller, W., *An Introduction to Probability Theory and its Applications, Volume II*, New York: John Wiley and Sons, 1966, p. 356.
11. Riordan, J., *Stochastic Service Systems*, New York: John Wiley and Sons, 1962, pp. 101–105.